A Fisher Information-based approach to improve labeling efficiency of neural network models in image classification

Joshua Caruso
Shreen Gul
Ardhendu Tripathy
Department of Computer Science
Funded by OURE Scholarship

Motivation

- Data is expensive/difficult to obtain, labeling is costly
- [1] Batch Active learning via Information maTrices (BAIT) reduces label cost with efficient data use.
- Training large models is computationally expensive
- [2] Fisher-Induced Sparse uncHanging Masks (FISH) train an important subset of the model.

Goals and Approach

- Make efficient use of limited data to reduct labeling cost
- Combine methods to expand BAIT sampling by applying [2] into [1]

Problem Setup

Image Classification

- We use the CIFAR-10 dataset of 50K images and ResNet-18 neural network

Active labeling

- The focus of active learning is on a function, Query
- In each epoch we choose a subset of data to label

BAIT

- Fisher information is defined:

$$I(x;\theta) = \mathbb{E}_{Y\sim p(y|x,\theta)}[\nabla^2 \log p(Y \mid x,\theta)]$$

- The next sample, x, is selected according to,

$$\arg\max_x \operatorname{tr}(V_x^\top M_i^{-1} I(\theta_t^L) M_i^{-1} V_x A^{-1})$$

where Mi is the fisher of selected samples, I is the fisher over last layer theta, Vx is a matrix of gradients and,

$$A = I + V_x^\top M_i^{-1} V_x$$

FISH

- Important weights are selected by largest fisher values calculated:

$$\widehat{F}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{Y\sim p(y|x,\theta)}[(\nabla \log p(Y \mid x,\theta))^2]$$

Chunksize

- Handle limited GPU memory and use multiple GPUs

```
for x_chunk in dataloader:
    trace[x_chunk] = diagonal(
    x_chunk @ Minv @ F @ Minv @ x_chunk.T,
    dim1=-2, dim2=-1).sum(-1)
```

Outcomes and Findings

Does chunksize affect calculation completion time?

- An almost periodic pattern is observed with decreasing amplitude

Do important weights change across rounds?

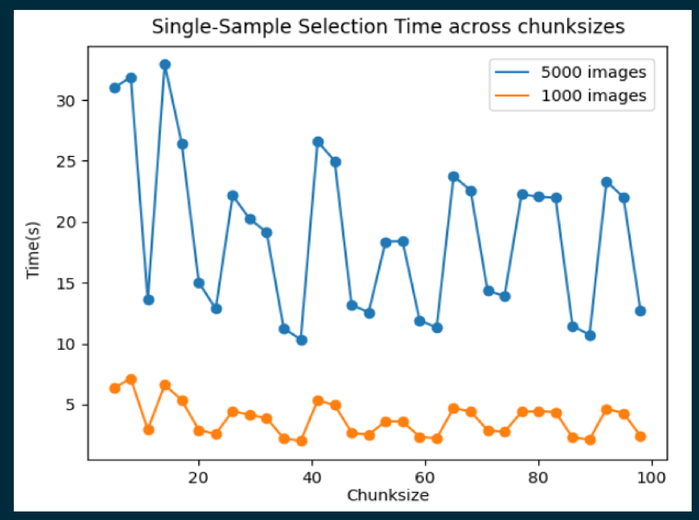
- Portion of a layer considered important exhibits a steep fall to nearly zero approximately one-third of the way into the model

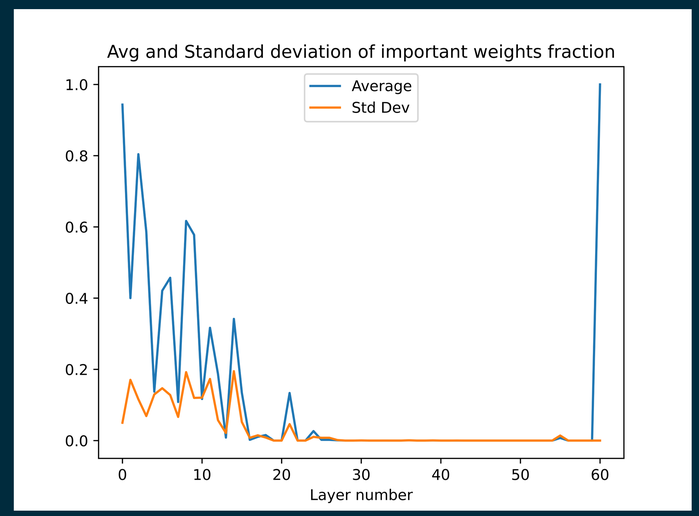
Does FISH mask result in improved performance of BAIT?

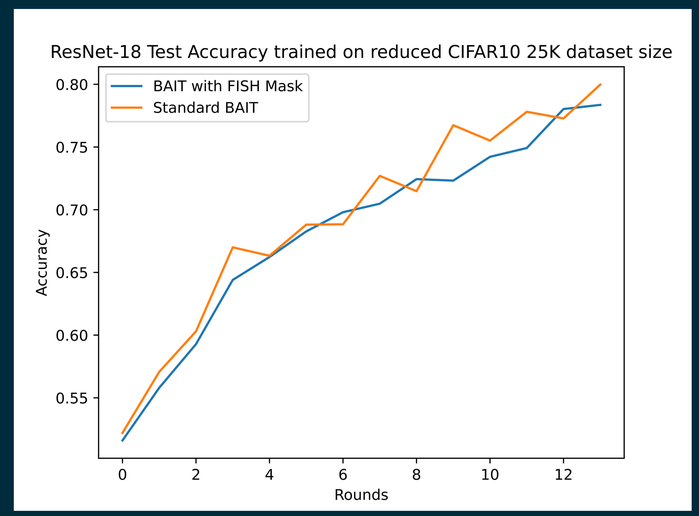
- Our combined approach currently produces results that are either as good as or insignificantly worse than standalone BAIT

Future Questions

- Does our combined approach see improvement over randomly selected theta?
- Does our approach see improvement over active learning with randomly selected samples?







[1] Ash, J., Goel, S., Krishnamurthy, A., & Kakade, S. (2021). Gone fishing: Neural active learning with fisher embeddings. Advances in Neural Information Processing Systems, 34, 8927-8939.

[2] Sung, Y. L., Nair, V., & Raffel, C. A. (2021). Training neural networks with fixed sparse masks. Advances in Neural Information Processing Systems, 34, 24193-24205.